MIDSEMESTER_BSE658_210615

2024-09-20

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

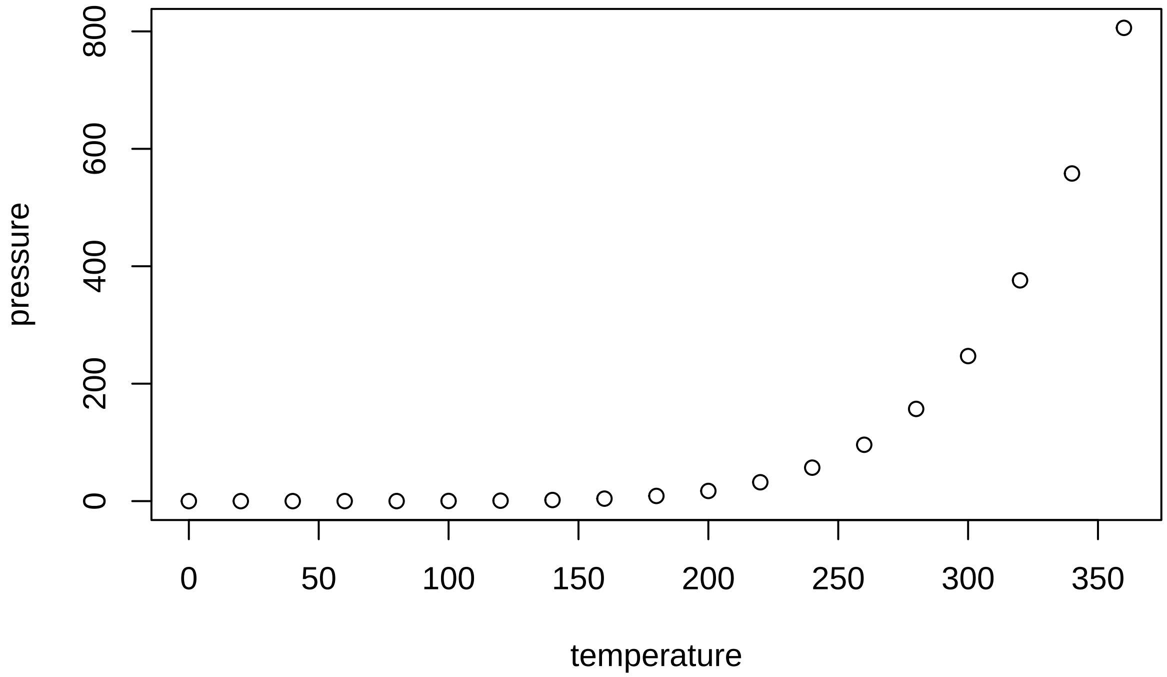
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(tibble)
library(ggplot2)
library(lsr)
library(psych)
```

```
##
## Attaching package: 'psych'
##
```

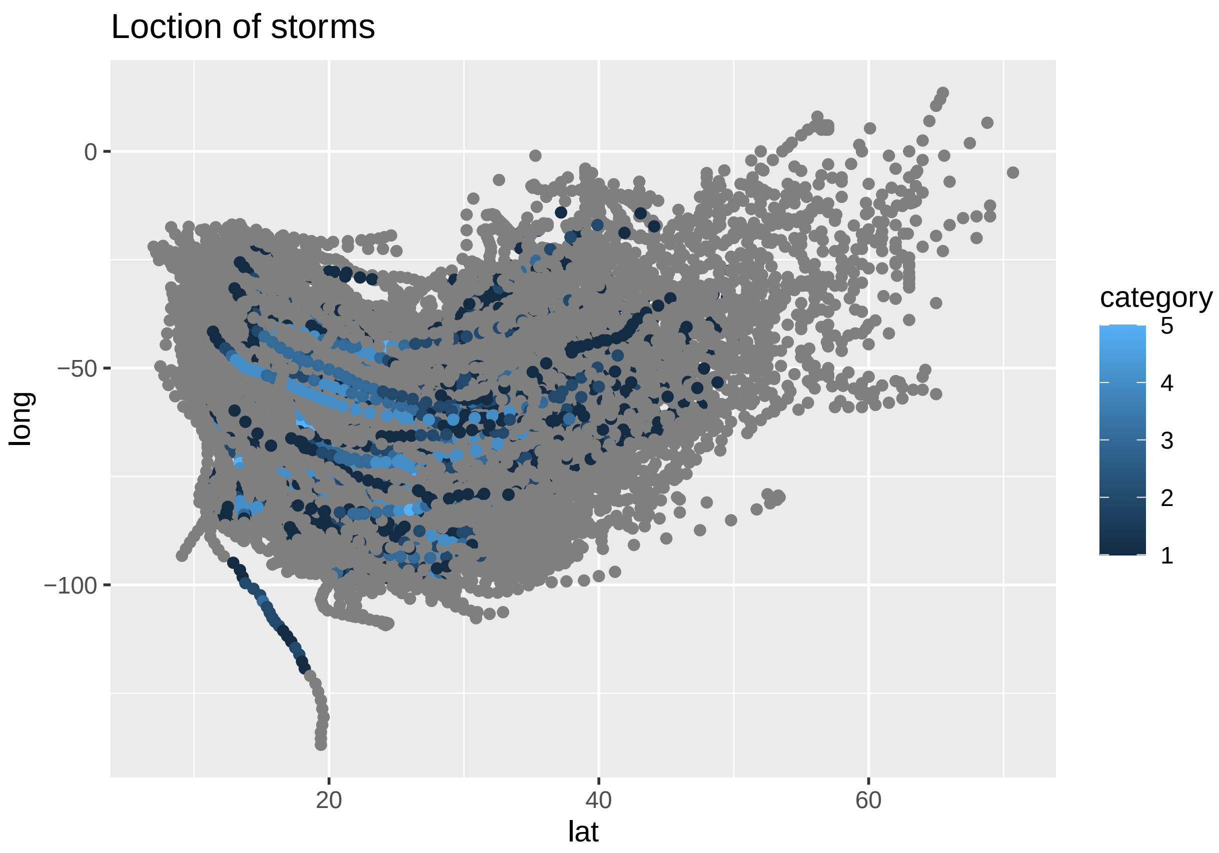
```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

Simpson's Paradox can lead to misleading conclusions when data is analyzed without considering confounding variables. In a medical study, the new treatment appeared less effective overall than the standard treatment. However, when stratified by condition severity, the new treatment was found to be more effective in each subgroup. This highlights the importance of accounting for confounding variables and using stratified analysis to avoid incorrect conclusions.

```r
data("storms")
head(storms)
```

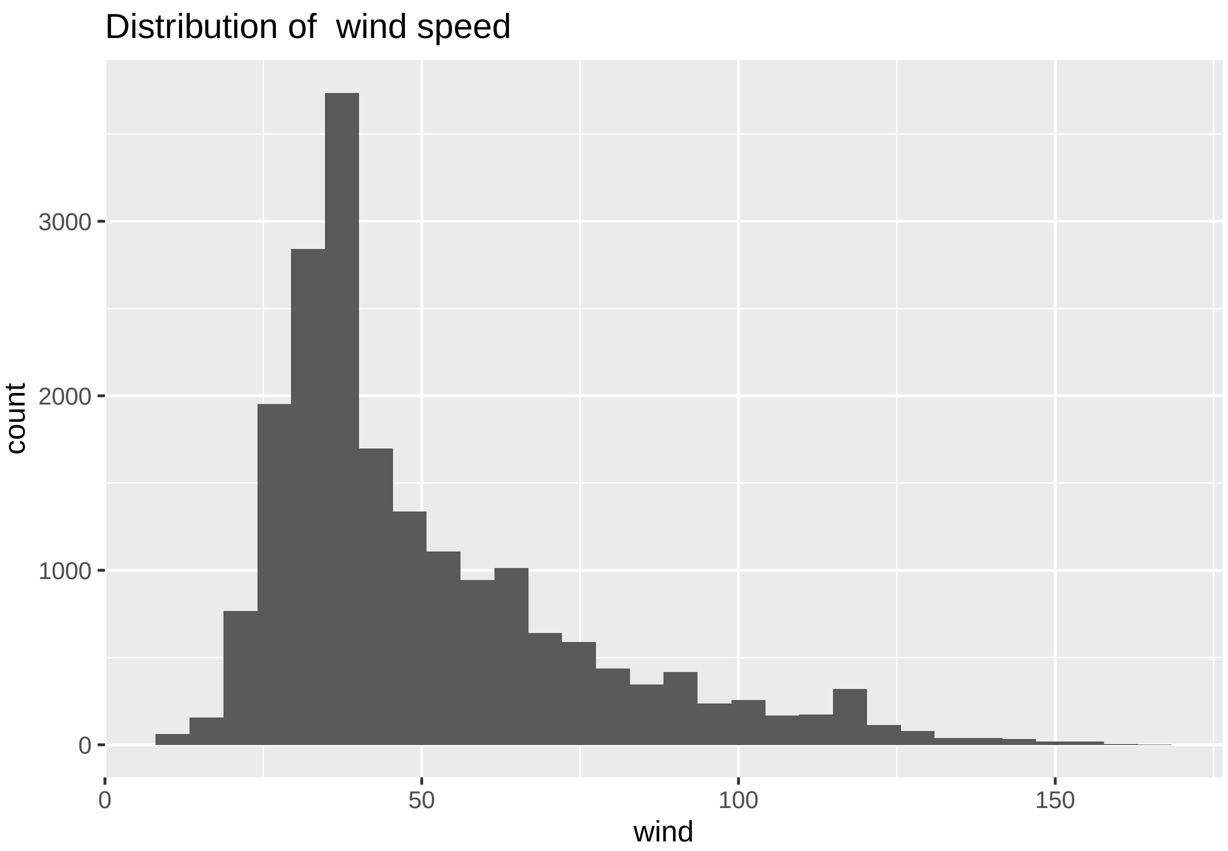
```
## # A tibble: 6 x 13
##   name   year month   day  hour   lat  long status       category  wind pressure
##   <chr> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <fct>           <dbl> <int>    <int>
## 1 Amy    1975     6    27     0  27.5 -79   tropical de~       NA    25     1013
## 2 Amy    1975     6    27     6  28.5 -79   tropical de~       NA    25     1013
## 3 Amy    1975     6    27    12  29.5 -79   tropical de~       NA    25     1013
## 4 Amy    1975     6    27    18  30.5 -79   tropical de~       NA    25     1013
## 5 Amy    1975     6    28     0  31.5 -78.8 tropical de~       NA    25     1012
## 6 Amy    1975     6    28     6  32.4 -78.7 tropical de~       NA    25     1012
## # i 2 more variables: tropicalstorm_force_diameter <int>,
## #   hurricane_force_diameter <int>
```

```r
ggplot(data=storms, aes(x = lat, y = long, color = category)) +
  geom_point() +
  labs(title = "Loction of storms")
```

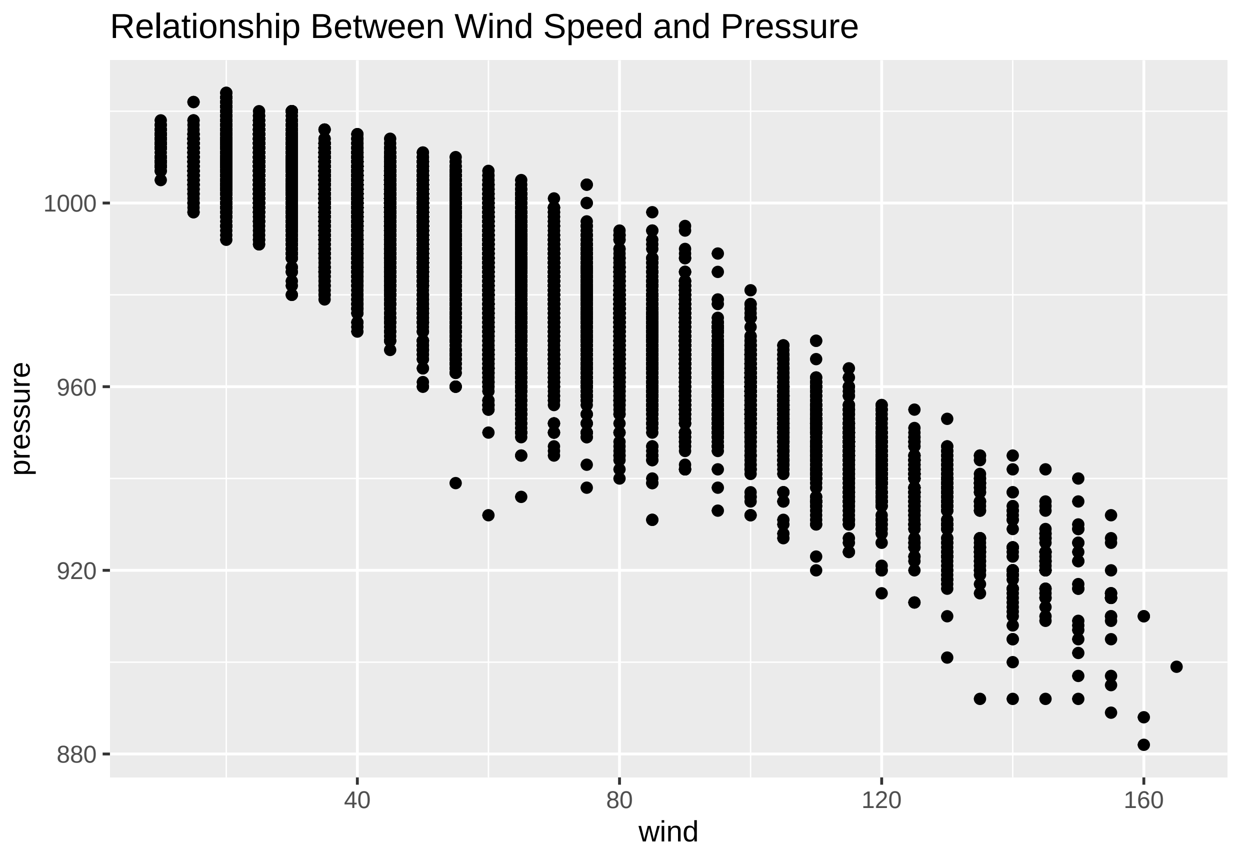


```
ggplot(data=storms, aes(x = wind)) +
  geom_histogram() +
  labs(title = "Distribution of  wind speed")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```r
ggplot(storms, aes(x = wind, y = pressure)) +
  geom_point() +
  labs(title = "Relationship Between Wind Speed and Pressure")
```



Conclusions : The wind speed is decreasing with pressure The highest wind speed had the lowest count

```
t_dist <- rnorm(10,10,5)
t_dist
```

```
##  [1]  8.672026 19.639815 10.861206 13.517987 10.055356 11.656415 10.553106
##  [8] 15.499024  5.547015  6.240963
```

```
ciMean(x=t_dist, conf =0.95)
```

```
##            2.5%    97.5%
## t_dist 8.207422 14.24116
```

```
mean(ciMean(x=t_dist, conf =0.95))
```

```
## [1] 11.22429
```

```
mean <- 10
sd <- 5
sample_size <- 10
samples <- rnorm(sample_size, mean = mean, sd = sd)
```

```r
ci_cimeans <- ciMean(samples)
t_value <- qt(0.975, df = sample_size - 1)
ci_tdist <- mean(samples) + c(-t_value, t_value) * sd(samples) / sqrt(sample_size)
cat("CI using cimeans:", ci_cimeans, "\n")
```

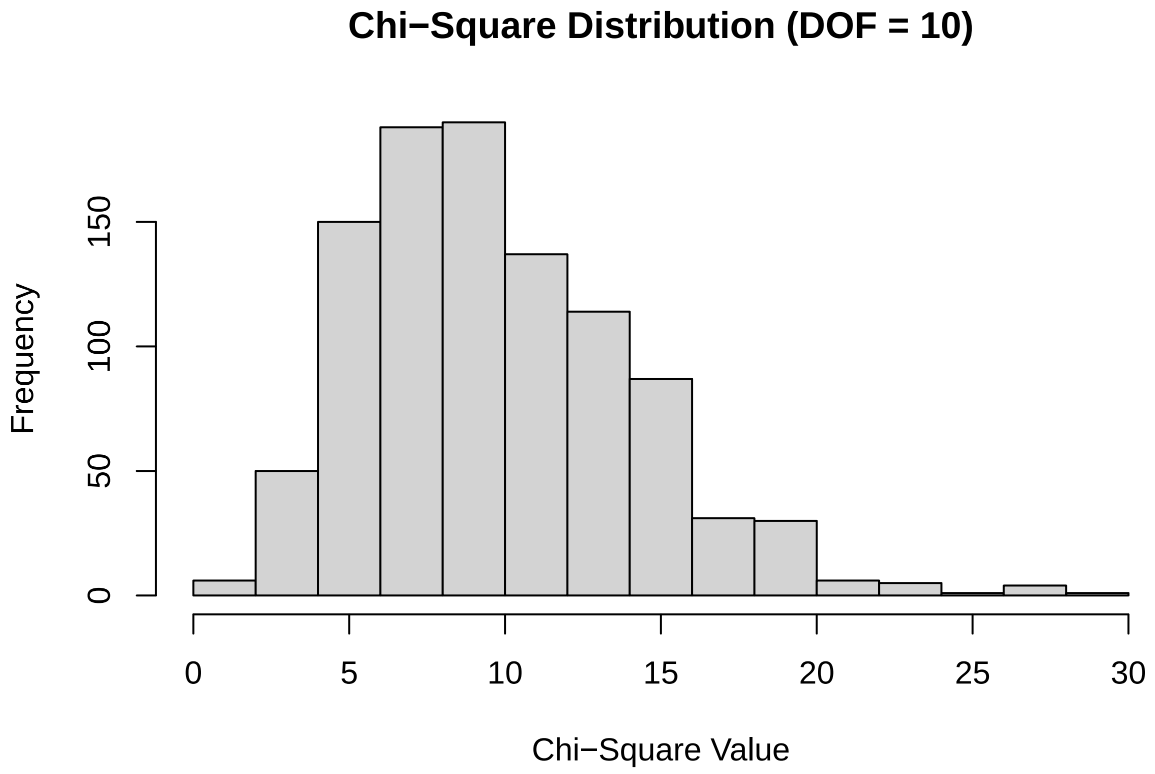
```
## CI using cimeans: 5.549533 11.82676
```

```r
cat("CI using t-distribution:", ci_tdist, "\n")
```

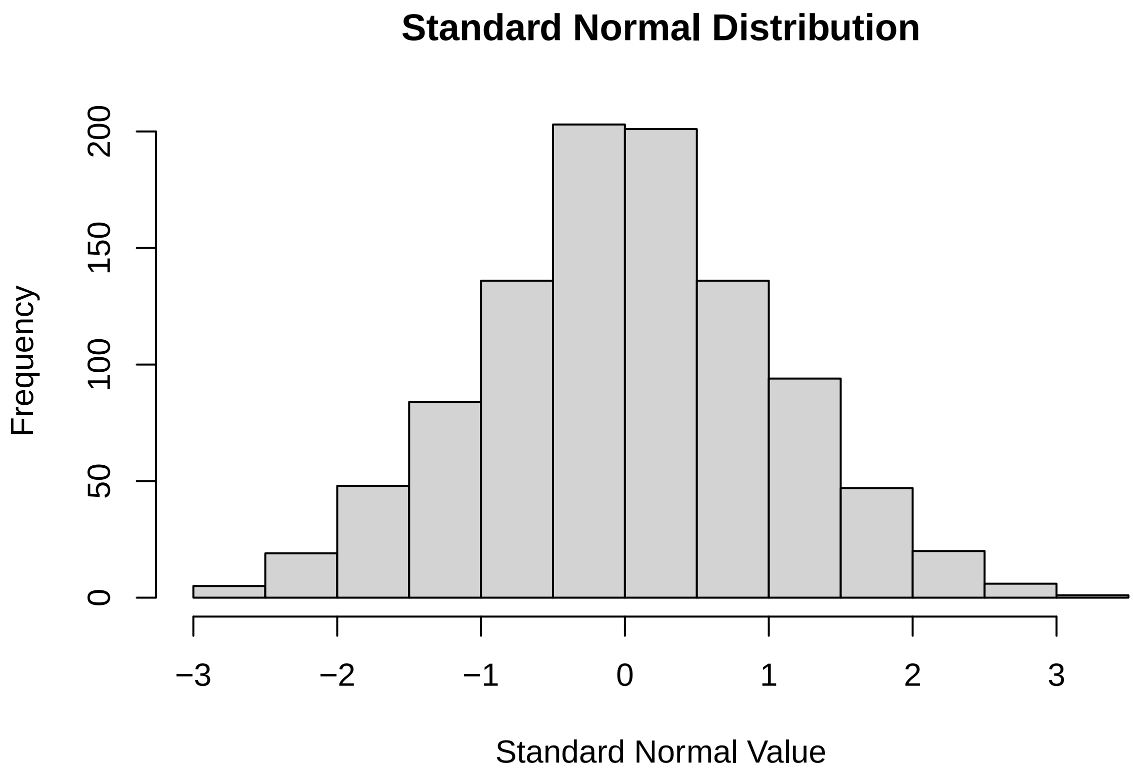
```
## CI using t-distribution: 5.549533 11.82676
```

A chi-square distribution is a probability distribution that describes the distribution of the sum of squared standard normal variables. It's often used in statistical hypothesis testing, particularly for goodness-of-fit tests and tests of independence. It is also used to as comparitive study for variances and distributions and independance.

```r
df <- 10

chi_sq_samples <- rchisq(1000, df)


hist(chi_sq_samples, main = "Chi-Square Distribution (DOF = 10)", xlab = "Chi-Square Value")
```



```r
normal_samples <- rnorm(1000, mean = 0, sd = 1)

hist(normal_samples, main = "Standard Normal Distribution", xlab = "Standard Normal Value")
```



The distribution got more skewed towards the right in case of chi square distribution compared to normal distribution

```r
x <- 5
sample_sizes <- c(10, 100, 1000)
for (n in sample_sizes) {
  samples <- rpois(n, x)
  sd_sample <- sd(samples)
  sem <- sd_sample / sqrt(n)
  cat("Sample size:", n, "SEM:", sem, "\n")
}
```

```
## Sample size: 10 SEM: 0.5333333
## Sample size: 100 SEM: 0.2073863
## Sample size: 1000 SEM: 0.07044062
```

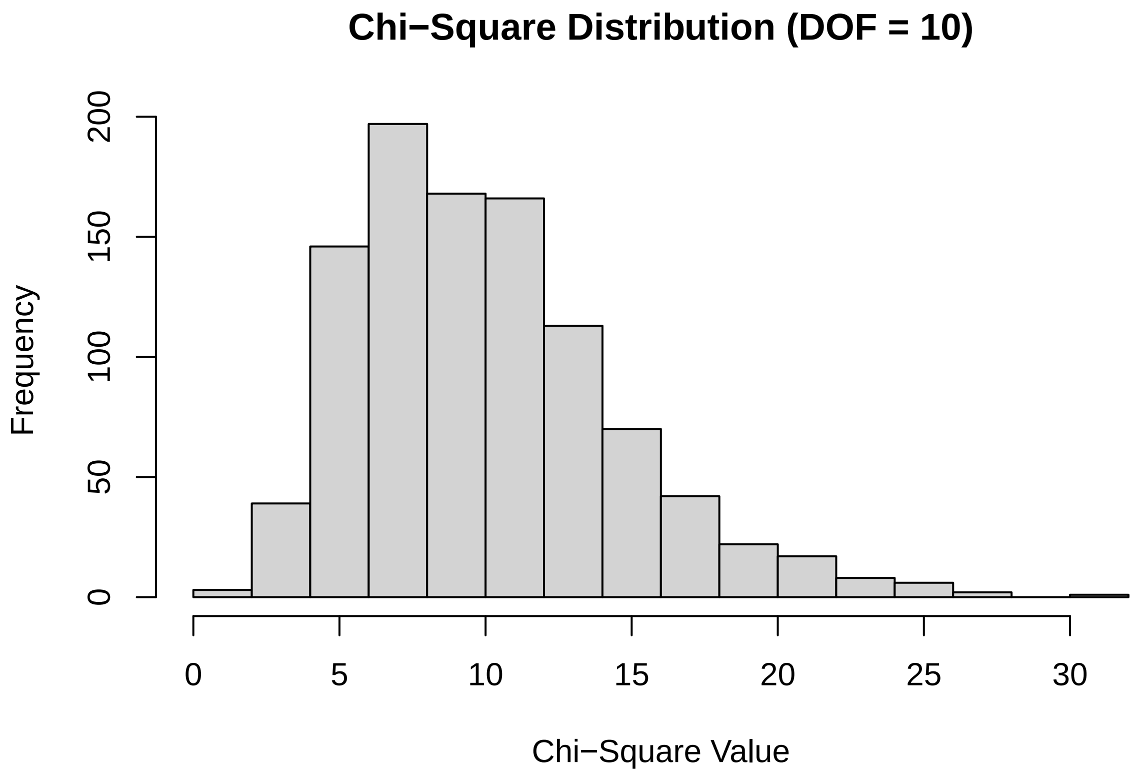
The standard error in mean is the error occured when we sample a population for a given number of times and calculate its mean and then plot a distribution out of it . The mean of that sample is the mean our sample mean. The variation or spread around it is the sample mean error . It is different from sample mean as the sample mean is the actual of randomly drawn out of sample, which may not even be close to true population mean. The population mean is the true mean that we get. When we increase the sample size the distribution of our mean becomes more and more like a normal distribution with mean equal to true population. We can clearly see that sample error is changing when the sample size is increasing showing the the graphs is getting towards more and more the true mean and lesser skewness.

When to Use:

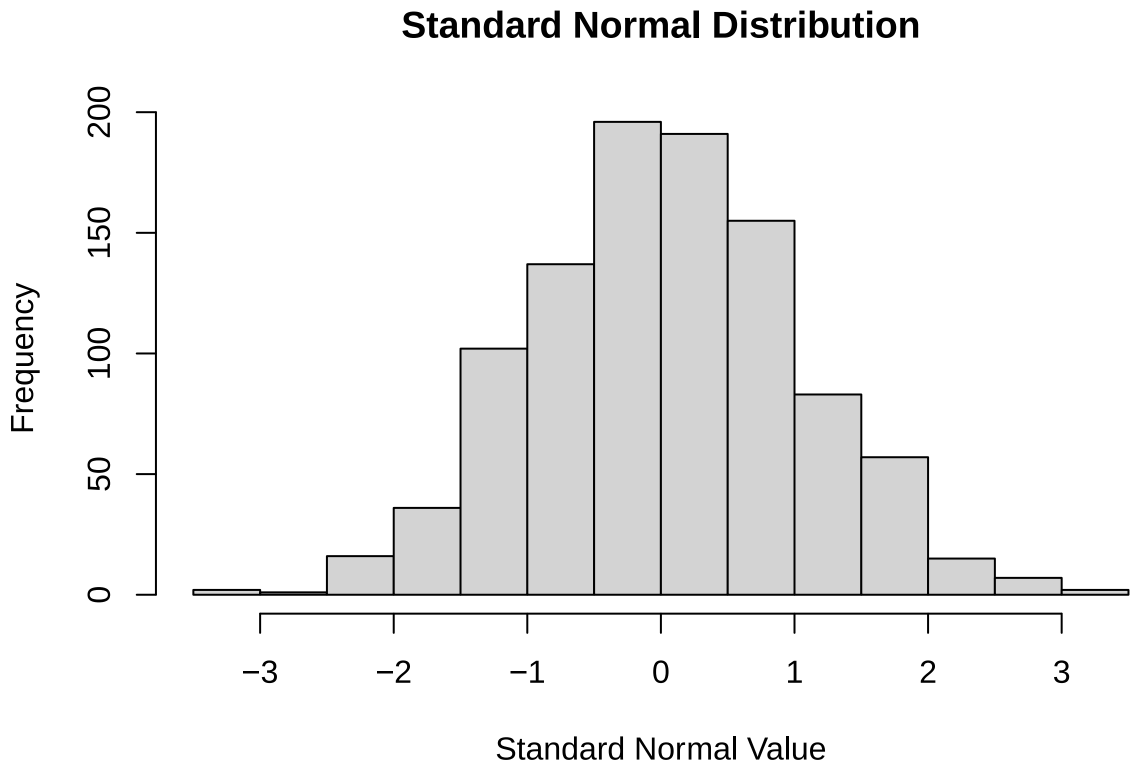
The chi-square distribution is commonly used in:

Goodness-of-Fit Tests: To assess how well observed data fits a theoretical distribution (e.g., comparing observed and expected frequencies in a contingency table). Tests of Independence: To determine if two categorical variables are independent (e.g., testing if there's a relationship between gender and voting preference). Hypothesis Testing: For various statistical tests, such as the chi-square test for variance and the chi-square test for independence.

```
df <- 10

chi_sq_samples <- rchisq(1000, df)


hist(chi_sq_samples, main = "Chi-Square Distribution (DOF = 10)", xlab = "Chi-Square Value")
```



```
normal_samples <- rnorm(1000, mean = 0, sd = 1)

hist(normal_samples, main = "Standard Normal Distribution", xlab = "Standard Normal Value")
```



1. Define Hypothesis: Null Hypothesis (H1) The new drug is not more effective than the old drug in reducing cholesterol levels. Alternative Hypothesis (H2): The new drug is more effective than the old drug in reducing cholesterol levels.

2. Collect Data:

Randomly assign participants to two groups: one receiving the new drug and the other receiving the old drug (or a placebo). Measure cholesterol levels before and after treatment for both groups.

3. Analyze Data:

Statistical Test: Choose a suitable statistical test, such as a paired t-test , based on the data distribution if it is normal or non normal. Calculate Test Statistic: Compute the test statistic using the collected data. 1. Calculate the Confidence Interval: Determine the desired confidence level (e.g., 95%). Calculate the appropriate confidence interval using the sample data and the chosen statistical method (e.g., t-interval for means, z-interval for proportions).

2. Interpret the Confidence Interval: If the confidence interval does not include the null value (e.g., 0 for a difference in means), it suggests that the effect is statistically significant at the chosen confidence level. The wider the confidence interval, the greater the uncertainty about the true population parameter.

Comparing Cholesterol Levels

Suppose we want to compare the mean cholesterol levels between two groups: those receiving the new drug and those receiving the old drug.

1. Calculate Confidence Intervals: Calculate 95% confidence intervals for the mean cholesterol levels in each group.

2. Compare Confidence Intervals: If the confidence intervals for the two groups do not overlap, it suggests that the difference between the means is statistically significant at a 95% confidence level. If the confidence intervals overlap, it indicates that the difference between the means may not be statistically significant.

3. Report Findings: Summarize Results: Present the statistical analysis, and findings. Draw Conclusions: Based on the evidence, conclude whether the new drug is more effective than the old drug in reducing cholesterol levels..

Ans